3.3 Non-Linear Regression

Many relationships between two variables follow patterns that are not linear. For example, square-law, exponential, and logarithmic relationships often appear in the natural sciences. **Non-linear regression** is an analytical technique for finding a curve of best fit for data from such relationships. The equation for this curve can then be used to model the relationship between the two variables.

As you might expect, the calculations for curves are more complicated than those for straight lines. Graphing calculators have built-in regression functions for a variety of curves, as do some spreadsheets and statistical programs. Once you enter the data and specify the type of curve, these technologies can automatically find the best-fit curve of that type. They can also calculate the coefficient of determination, r^2, which is a useful measure of how closely a curve fits the data.

INVESTIGATE & INQUIRE: Bacterial Growth

A laboratory technician monitors the growth of a bacterial culture by scanning it every hour and estimating the number of bacteria. The initial population is unknown.

Time (h)	0	1	2	3	4	5	6	7
Population	?	10	21	43	82	168	320	475

1. a) Create a scatter plot and classify the linear correlation.
   b) Determine the correlation coefficient and the line of best fit.
   c) Add the line of best fit to your scatter plot. Do you think this line is a satisfactory model? Explain why or why not.
2. a) Use software or a graphing calculator to find a curve of best fit with a
      i) quadratic regression of the form $y = ax^2 + bx + c$
      ii) cubic regression of the form $y = ax^3 + bx^2 + cx + d$
   b) Graph these curves onto a scatter plot of the data.
   c) Record the equation and the coefficient of determination, r^2, for the curves.
   d) Use the equations to estimate the initial population of the bacterial culture. Do these estimates seem reasonable? Why or why not?

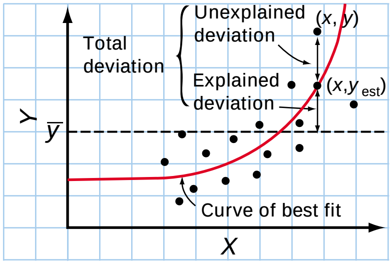
See Appendix B for details on using technology for non-linear regressions.

3. a) Perform an exponential regression on the data. Graph the curve of best fit and record its equation and coefficient of determination.
   b) Use this model to estimate the initial population.
   c) Do you think the exponential equation is a better model for the growth of the bacterial culture than the quadratic or cubic equations? Explain your reasoning.

Recall that Pearson's correlation coefficient, r, is a measure of the linearity of the data, so it can indicate only how closely a straight line fits the data. However, the **coefficient of determination, r^2**, is defined such that it applies to any type of regression curve.

$$r^2 = \frac{\text{variation in } y \text{ explained by variation in } x}{\text{total variation in } y}$$
$$= \frac{\Sigma(y_{\text{est}} - \overline{y})^2}{\Sigma(y - \overline{y})^2}$$

where $\overline{y}$ is the mean y value, y_{est} is the value estimated by the best-fit curve for a given value of x, and y is the actual observed value for a given value of x.



The total variation is the sum of the squares of the deviations for all of the individual data points.

The coefficient of determination can have values from 0 to 1. If the curve is a perfect fit, then y_{est} and y will be identical for each value of x. In this case, the variation in x accounts for all of the variation in y, so $r^2 = 1$. Conversely, if the curve is a poor fit, the total of $(y_{\text{est}} - \overline{y})^2$ will be much smaller than the total of $(y - \overline{y})^2$, since the variation in x will account for only a small part of the total variation in y. Therefore, r^2 will be close to 0. For any given type of regression, the curve of best fit will be the one that has the highest value for r^2.

For graphing calculators and Microsoft® Excel, the procedures for non-linear regression are almost identical to those for linear regression. At present, Corel® Quattro® Pro and Fathom™ do not have built-in functions for non-linear regression.

Exponential Regression

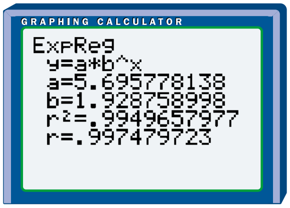
Exponential regressions produce equations with the form $y = ab^x$ or $y = ae^{kx}$, where $e = 2.718\ 28\ldots$, an irrational number commonly used as the base for exponents and logarithms. These two forms are equivalent, and it is straightforward to convert from one to the other.

Example 1 Exponential Regression

Generate an exponential regression for the bacterial culture in the investigation on page 184. Graph the curve of best fit and determine its equation and the coefficient of determination.

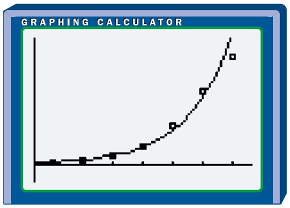
Solution 1 Using a Graphing Calculator

Use the ClrList command from the STAT EDIT menu to clear lists L1 and L2, and then enter the data. Set DiagnosticOn so that regression calculations will display the coefficient of determination. From the STAT CALC menu, select the non-linear regression function ExpReg. If you do not enter any list names, the calculator will use L1 and L2 by default.



```
ExpReg
 y=a*b^x
 a=5.695778138
 b=1.928758998
 r²=.9949657977
 r=.997479723
```

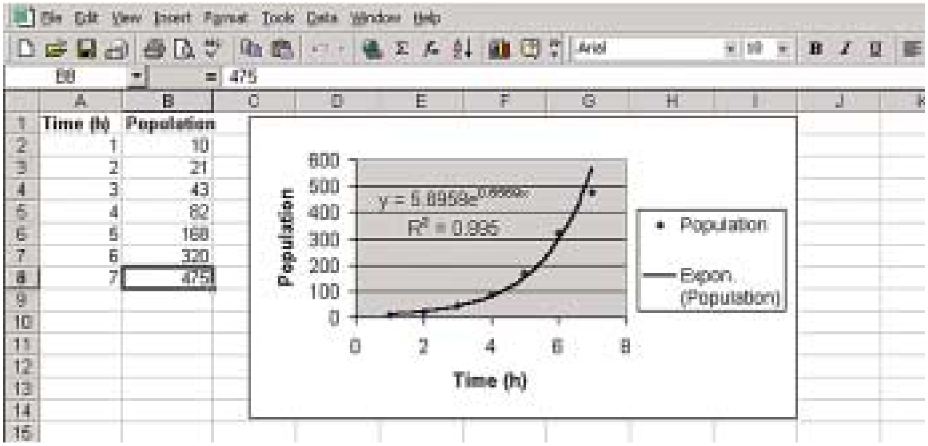
The equation for the curve of best fit is $y = 5.70(1.93)^x$, and the coefficient of determination is $r^2 = 0.995$. Store the equation as Y1. Use STAT PLOT to display a scatter plot of the data along with Y1. From the ZOOM menu, select 9:ZoomStat to adjust the window settings automatically.



Solution 2 Using a Spreadsheet

Enter the data into two columns. Next, highlight these columns and use the Chart feature to create an x-y scatter plot.

Select Chart/Add Trendline and then choose Expontenial regression. Then, select the curve that appears on your chart, and click Format/Selected Trendline/Options. Check the option boxes to display the equation and r^2.



B8 = 475

	Time (h)	Population
1	Time (h)	Population
2	1	10
3	2	21
4	3	43
5	4	82
6	5	168
7	6	320
8	7	475

The equation of the best-fit curve is $y = 5.7e^{0.66x}$ and the coefficient of determination is $r^2 = 0.995$. This equation appears different from the one found with the graphing calculator. In fact, the two forms are equivalent, since $e^{0.66} \doteq 1.93$.

Power and Polynomial Regression

In **power regressions**, the curve of best fit has an equation with the form $y = ax^b$.

Example 2 Power Regression

For a physics project, a group of students videotape a ball dropped from the top of a 4-m high ladder, which they have marked every 10 cm. During playback, they stop the videotape every tenth of a second and compile the following table for the distance the ball travelled.

Time (s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Distance (m)	0.05	0.2	0.4	0.8	1.2	1.7	2.4	3.1	3.9	4.9

a) Does a linear model fit the data well?

b) Use a power regression to find a curve of best fit for the data. Does the power-regression curve fit the data more closely than the linear model does?

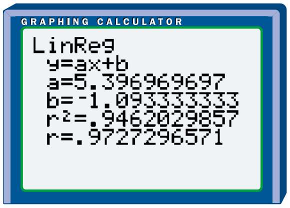
c) Use the equation for the regression curve to predict

i) how long the ball would take to fall 10 m

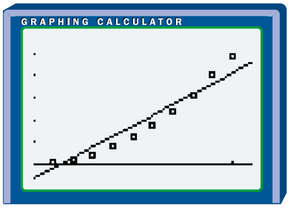
ii) how far the ball would fall in 5 s

Solution 1 Using a Graphing Calculator

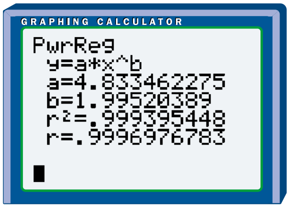
a) Although the linear correlation coefficient is 0.97, a scatter plot of the data shows a definite curved pattern. Since $b = -1.09$, the linear model predicts an initial position of about –1.1 m and clearly does not fit the first part of the data well. Also, the pattern in the scatter plot suggests the linear model could give inaccurate predictions for times beyond 1 s.



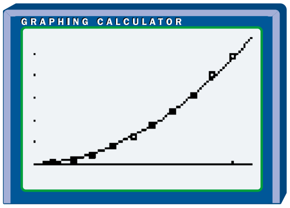
```
LinReg
 y=ax+b
 a=5.396969697
 b=-1.093333333
 r²=.9462029857
 r=.9727296571
```



b) From the STAT CALC menu, select the non-linear regression function PwrReg and then follow the same steps as in Example 1.



```
PwrReg
 y=a*x^b
 a=4.833462275
 b=1.99520389
 r²=.999395448
 r=.9996976783
```



The equation for the curve of best fit is $y = 4.83x^2$. The coefficient of determination and a graph on the scatter plot show that the quadratic curve is almost a perfect fit to the data.

c) Substitute the known values into the equation for the quadratic curve of best fit:

i)
$$\begin{aligned} 10 &= 4.83x^2 \\ x^2 &= \frac{10}{4.83} \\ x &= \sqrt{\frac{10}{4.83}} \\ &= 1.4 \end{aligned}$$

ii)
$$\begin{aligned} y &= 4.83(5)^2 \\ &= 4.83(25) \\ &= 121 \end{aligned}$$

The quadratic model predicts that

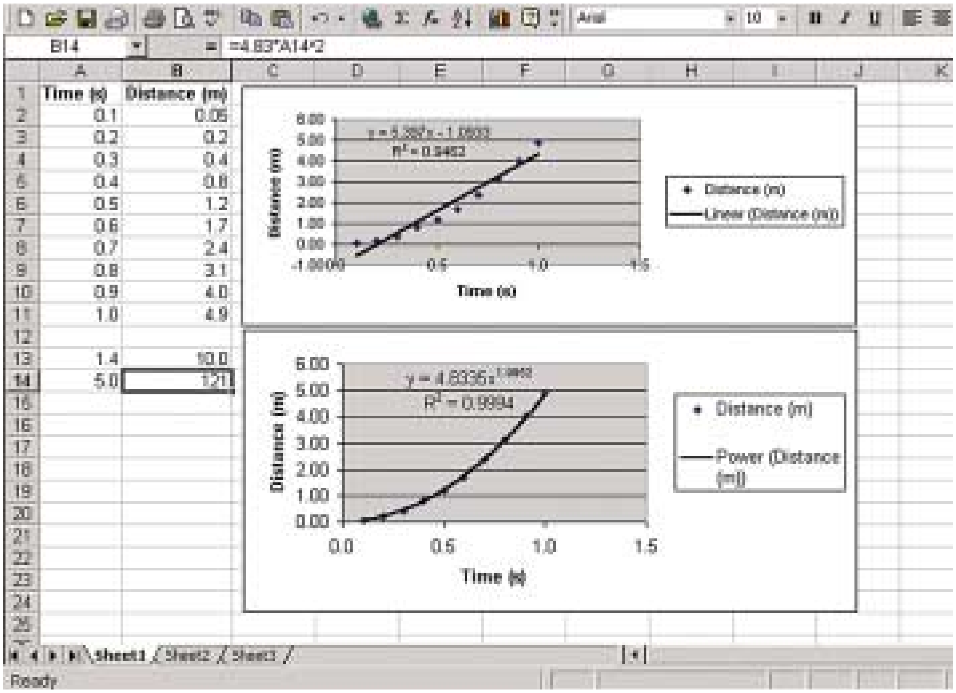
i) the ball would take approximately 1.4 s to fall 10 m

ii) the ball would fall 121 m in 5 s

Solution 2 Using a Spreadsheet

a) As in Solution 1, the scatter plot shows that a curve might be a better model.

b) Use the **Chart feature** as in Example 1, but select Power when adding the trend line.



B14 =4.83*A14^2

	A	B
1	Time (s)	Distance (m)
2	0.1	0.05
3	0.2	0.2
4	0.3	0.4
5	0.4	0.8
6	0.5	1.2
7	0.6	1.7
8	0.7	2.4
9	0.8	3.1
10	0.9	4.0
11	1.0	4.9
12		
13	1.4	10.0
14	5.0	121

The equation for the curve of best fit is $y = 4.83x^2$. The graph and the value for r^2 show that the quadratic curve is almost a perfect fit to the data.

c) Use the equation for the curve of best fit to enter formulas for the two values you want to predict, as shown in cells A13 and B14 in the screen above.

Example 3 Polynomial Regression

Suppose that the laboratory technician takes further measurements of the bacterial culture in Example 1.

Time (h)	8	9	10	11	12	13	14
Population	630	775	830	980	1105	1215	1410

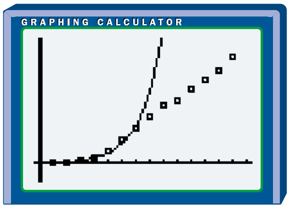
a) Discuss the effectiveness of the exponential model from Example 1 for the new data.

b) Find a new exponential curve of best fit.

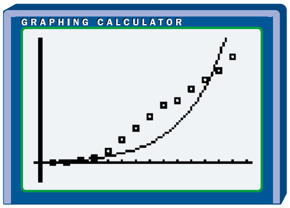
c) Find a better curve of best fit. Comment on the effectiveness of the new model.

Solution

a) If you add the new data to the scatter plot, you will see that the exponential curve determined earlier, $y = 5.7(1.9)^x$, is no longer a good fit.

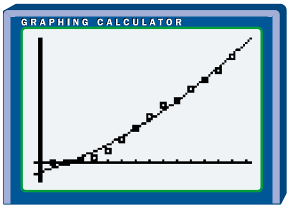


b) If you perform a new exponential regression on all 14 data points, you obtain the equation $y = 18(1.4)^x$ with a coefficient of determination of $r^2 = 0.88$. From the graph, you can see that this curve is not a particularly good fit either.



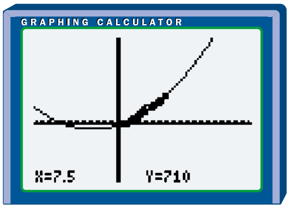
Because of the wide range of non-linear regression options, you can insist on a fairly high value of r^2 when searching for a curve of best fit to model the data.

c) If you perform a quadratic regression, you get a much better fit with the equation $y = 4.0x^2 + 55x - 122$ and a coefficient of determination of $r^2 = 0.986$.



This quadratic model will probably serve well for interpolating between most of the data shown, but may not be accurate for times before 3 h and after 14 h. At some point between 2 h and 3 h, the curve intersects the x-axis, indicating a negative population prior to this time. Clearly the quadratic model is not accurate in this range.

Similarly, if you zoom out, you will notice a problem beyond 14 h. The rate of change of the quadratic curve continues to increase after 14 h, but the trend of the data does not suggest such an increase. In fact, from 7 h to 14 h the trend appears quite linear.



It is important to recognize the limitations of regression curves. One interesting property of polynomial regressions is that for a set of n data points, a polynomial function of degree $n - 1$ can be produced which perfectly fits the data, that is, with $r^2 = 1$.

For example, you can determine the equation for a line (a first-degree polynomial) with two points and the equation for a quadratic (a second-degree polynomial) with three points. However, these polynomials are not always the best models for the data. Often, these curves can give inaccurate predictions when extrapolated.

Sometimes, you can find that several different types of curves fit closely to a set of data. Extrapolating to an initial or final state may help determine which model is the most suitable. Also, the mathematical model should show a logical relationship between the variables.

Project Prep

Non-linear models may be useful when you are analysing two-variable data in your statistics project.

Key Concepts

- Some relationships between two variables can be modelled using non-linear regressions such as quadratic, cubic, power, polynomial, and exponential curves.
- The coefficient of determination, r^2, is a measure of how well a regression curve fits a set of data.
- Sometimes more than one type of regression curve can provide a good fit for data. To be an effective model, however, the curve must be useful for extrapolating beyond the data.

Communicate Your Understanding

1. A data set for two variables has a linear correlation coefficient of 0.23. Does this value preclude a strong correlation between the variables? Explain why or why not.
2. A best-fit curve for a set of data has a coefficient of determination of $r^2 = 0.76$. Describe some techniques you can use to improve the model.

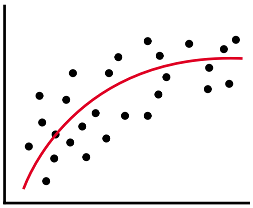
Practise

A

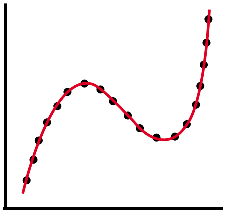
1. Match each of the following coefficients of determination with one of the diagrams below.

a) 0 **b)** 0.5 **c)** 0.9 **d)** 1

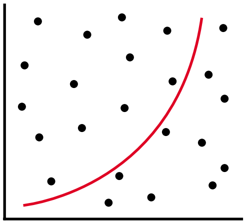
i)



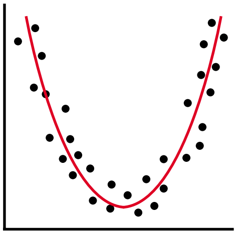
ii)



iii)



iv)



2. For each set of data use software or a graphing calculator to find the equation and coefficient of determination for a curve of best fit.

a)

x	y
−2.8	0.6
−3.5	−5.8
−2	3
−1	6
0.2	4
1	1
−1.5	5
1.4	−3.1
0.7	3
−0.3	6.1
−3.3	−3.1
−4	−7
2	−5.7

b)

x	y
−2.7	1.6
−3.5	−3
−2.2	3
−0.5	−0.5
0	1.3
0.6	4.7
−1.8	1.7
−3.8	−7
−1.3	0.6
0.8	7
0.5	2.7
−1	1.5
−3	−1.1

c)

x	y
1.1	2.5
3.5	11
2.8	8.6
2.3	7
0	1
3.8	14
1.4	4.2
−4	0.2
−1.3	0.6
3	12
4.1	17
2.2	5
−2.7	0.4

Apply, Solve, Communicate

B

3. The heights of a stand of pine trees were measured along with the area under the cone formed by their branches.

Height (m)	Area (m^2)
2.0	5.9
1.5	3.4
1.8	4.8
2.4	8.6
2.2	7.3
1.2	2.1
1.8	4.9
3.1	14.4

a) Create a scatter plot of these data.

b) Determine the correlation coefficient and the equation of the line of best fit.

c) Use a power regression to calculate a coefficient of determination and an equation for a curve of best fit.

d) Which model do you think is more accurate? Explain why.

e) Use the more accurate model to predict

i) the area under a tree whose height is 2.7 m

ii) the height of a tree whose area is 30 m^2

f) Suggest a reason why the height and circumference of a tree might be related in the way that the model in part d) suggests.

4. **Application** The biologist Max Kleiber (1893–1976) pioneered research on the metabolisms of animals. In 1932, he determined the relationship between an animal's mass and its energy requirements or basal metabolic rate (BMR). Here are data for eight animals.

Animal	Mass (kg)	BMR (kJ/day)
Frog	0.018	0.050
Squirrel	0.90	1.0
Cat	3.0	2.6
Monkey	7.0	4.0
Baboon	30	14
Human	60	25
Dolphin	160	44
Camel	530	116

a) Create a scatter plot and explain why Kleiber thought a power-regression curve would fit the data.

b) Use a power regression to find the equation of the curve of best fit. Can you rewrite the equation so that it has exponents that are whole numbers? Do so, if possible, or explain why not.

c) Is this power equation a good mathematical model for the relationship between an animal's mass and its basal metabolic rate? Explain why or why not.

d) Use the equation of the curve of best fit to predict the basal metabolic rate of

i) a 15-kg dog

ii) a 2-tonne whale

5. **Application** As a sample of a radioactive element decays into more stable elements, the amount of radiation it gives off decreases. The level of radiation can be used to estimate how much of the original element remains. Here are measurements for a sample of radium-227.

Time (h)	Radiation Level (%)
0	100
1	37
2	14
3	5.0
4	1.8
5	0.7
6	0.3

a) Create a scatter plot for these data.

b) Use an exponential regression to find the equation for the curve of best fit.

c) Is this equation a good model for the radioactive decay of this element? Explain why or why not.

d) A half-life is the time it takes for half of the sample to decay. Use the regression equation to estimate the half-life of radium-227.

6. Chapter Problem **a)** Create a time-series graph for the mean starting salary of the graduates who find jobs. Describe the pattern that you see.

b) Use non-linear regression to construct a curve of best fit for the data. Record the equation of the curve and the coefficient of determination.

c) Comment on whether this equation is a good model for the graduates' starting salaries.

7. An engineer testing the transmitter for a new radio station measures the radiated power at various distances from the transmitter. The engineer's readings are in microwatts per square metre.

Distance (km)	Power Level ($\mu W/m^2$)
2.0	510
5.0	78
8.0	32
10.0	19
12.0	14
15.0	9
20.0	5

a) Find an equation for a curve of best fit for these data that has a coefficient of determination of at least 0.98.

b) Use the equation for this curve of best fit to estimate the power level at a distance of

i) 1.0 km from the transmitter

ii) 4.0 km from the transmitter

iii) 50.0 km from the transmitter

8. **Communication** *Logistic* curves are often a good model for population growth. These curves have equations with the form $y = \dfrac{c}{1 + ae^{-bx}}$, where a, b, and c are constants. Consider the following data for the bacterial culture in Example 1:

Time (h)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Population	?	10	21	43	82	168	320	475	630	775	830	980	1105	1215	1410	1490	1550	1575	1590	1600	1600

a) Use software or a graphing calculator to find the equation and coefficient of determination for the logistic curve that best fits the data for the bacteria population from 1 to 20 h.

b) Graph this curve on a scatter plot of the data.

c) How well does this curve appear to fit the entire data set? Describe the shape of the curve.

d) Write a brief paragraph to explain why you think a bacterial population may exhibit this type of growth pattern.

9. **Inquiry/Problem Solving** The following table shows the estimated population of a crop-destroying insect.

Year	Population (billions)
1995	100
1996	130
1997	170
1998	220
1999	285
2000	375
2001	490

a) Determine an exponential curve of best fit for the population data.

b) Suppose that 100 million of an arachnid that preys on the insect are imported from overseas in 1995. Assuming the arachnid population doubles every year, estimate when it would equal 10% of the insect population.

c) What further information would you need in order to estimate the population of the crop-destroying insect once the arachnids have been introduced?

d) Write an expression for the size of this population.

C

10. Use technology to calculate the coefficient of determination for two of the linear regression examples in section 3.2. Is there any relationship between these coefficients of determination and the linear correlation coefficients for these examples?

11. **Inquiry/Problem Solving** Use a software program, such as Microsoft® Excel, to analyse these two sets of data:

Data Set A		Data Set B	
x	*y*	*x*	*y*
2	5	2	6
4	7	4	5
6	2	7	−4
8	5	9	1
		12	2

a) For each set of data,

i) determine the degree of polynomial regression that will generate a perfectly fit regression curve

ii) perform the polynomial regression and record the value of r^2 and the equation of the regression curve

b) Assess the effectiveness of the best-fit polynomial curve as a model for the trend of the set of data.

c) For data set B,

i) explain why the best-fit polynomial curve is an unsatisfactory model

ii) generate a better model and record the value of r^2 and the equation of your new best-fit curve

iii) explain why this curve is a better model than the polynomial curve found in part a)